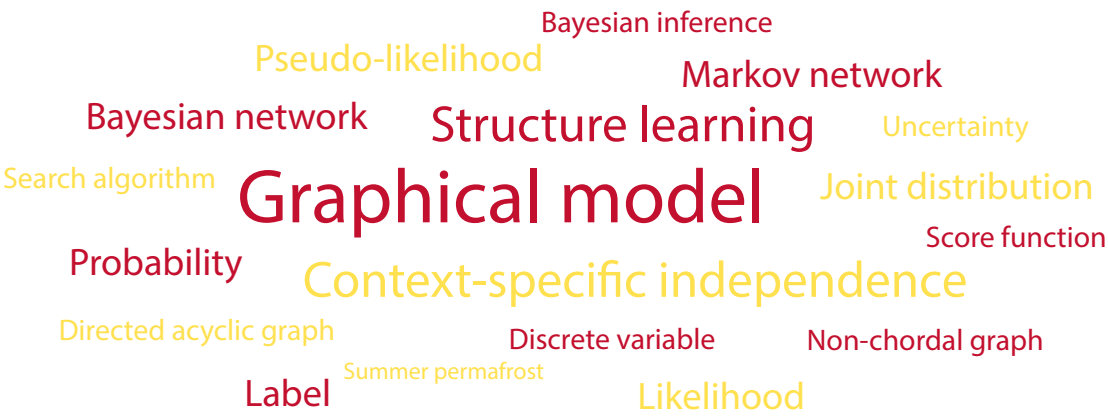
Åbo Akademi University

Johan Pensar

Structure Learning of Context-Specific Graphical Models



Structure Learning of Context-Specific Graphical Models

Johan Pensar

PhD Thesis in Statistics
Mathematics and Statistics
Faculty of Science and Engineering
Åbo Akademi University

Åbo, Finland, 2016

Supervisor

Professor Jukka Corander,
Department of Mathematics and Statistics,
University of Helsinki,
Helsinki, Finland

Reviewers

Assistant Professor Mikko Koivisto,
Department of Computer Science,
University of Helsinki,
Helsinki, Finland

Professor Dan Geiger,
Department of Computer Science,
Technion - Israel Institute of Technology,
Haifa, Israel

Opponent

Professor Dan Geiger,
Department of Computer Science,
Technion - Israel Institute of Technology,
Haifa, Israel

ISBN 978-952-12-3412-5

Painosalama Oy
Åbo, Finland, 2016

Preface

This thesis concludes the research I have conducted during the years 2012-2016 at the subject of Mathematics and Statistics at Åbo Akademi University. During this time, I have not only grown as a researcher, but also as a person. This would not have been possible without the endless support from the people around me. For this, I am truly grateful.

First of all, I would like to thank my supervisor Jukka Corander whose help and support I have always been able to count on, although we have worked in different cities. There are few persons with the same amount of positivity and never-ending enthusiasm, whether is about tackling a research problem or building a "stone bridge" for crossing a melt-water stream in the Swiss alps.

I would also like to direct a special thanks to my colleague and friend Henrik Nyman, the second member of team BAPS, Åbo division. It has been a pleasure sharing office and discussing various research and non-research related issues. At this point, it is also in place to thank the friendly staff at Hotel Cepina, the venue of Summer Permafrost 2012-2015, where many of the ideas leading to this thesis were generated and developed.

I would like to thank everybody at the subject of Mathematics and Statistics for the excellent work atmosphere. In particular, I would like to thank Göran Högnäs and Paavo Salminen, who have helped me with all kinds of practical matters during my PhD studies.

I thank all co-authors of the included articles for their contributions. I further thank Mikko Koivisto for reviewing my thesis and Dan Geiger for taking the time to both review and act as opponent.

For their financial support, I gratefully acknowledge the Magnus Ehrnrooth foundation, the Finnish Doctoral Programme in Stochastics and Statistics (FDPSS), the Center of Excellence in Optimization and Systems Engineering at Åbo Akademi University, and Åbo Akademi University.

Finally, and most importantly, a big thanks goes to my family and friends. It is safe to say that I would not be where I am today without the support and care from my family. Of course, I also warmly thank my ray of sunshine Heidi whom I can always rely on to brighten my day.

Åbo, May 2016

Johan Pensar

Abstract

The ultimate problem considered in this thesis is modeling a high-dimensional joint distribution over a set of discrete variables. For this purpose, we consider classes of context-specific graphical models and the main emphasis is on learning the structure of such models from data. Traditional graphical models compactly represent a joint distribution through a factorization justified by statements of conditional independence which are encoded by a graph structure. Context-specific independence is a natural generalization of conditional independence that only holds in a certain context, specified by the conditioning variables. We introduce context-specific generalizations of both Bayesian networks and Markov networks by including statements of context-specific independence which can be encoded as a part of the model structures. For the purpose of learning context-specific model structures from data, we derive score functions, based on results from Bayesian statistics, by which the plausibility of a structure is assessed. To identify high-scoring structures, we construct stochastic and deterministic search algorithms designed to exploit the structural decomposition of our score functions. Numerical experiments on synthetic and real-world data show that the increased flexibility of context-specific structures can more accurately emulate the dependence structure among the variables and thereby improve the predictive accuracy of the models.

Sammanfattning

Det grundläggande problemet som behandlas i denna avhandling är modellering av en högdimensionell simultan fördelning över en mängd diskreta variabler. För detta ändamål undersöker vi klasser av kontextspecifika grafiska modeller och vi fokuserar på inlärningen av modellstrukturen från data. Traditionella grafiska modeller utgör en kompakt representation av en simultan fördelning genom att faktorisera fördelningen enligt en graf som återspeglar antaganden om betingat oberoende. Betingat oberoende har en naturlig generalisering i kontextspecifikt oberoende som endast håller i en viss kontext som bestäms av de betingande variablerna. Vi introducerar kontextspecifika generaliseringar av både Baysianska nätverk och Markovnätverk genom att inkludera kontextspecifika oberoenden som en del av modellstrukturerna. För inlärningen av kontextspecifika modellstrukturer från data använder vi oss av resultat från Bayesiansk statistik för att härleda målfunktioner som bedömer trovärdigheten av en viss struktur. För att identifiera strukturer med hög trovärdighet används deterministiska och stokastiska sökalgoritmer som är designade att utnyttja strukturen i målfunktionernas faktorisering. Numeriska experiment baserade på syntetiska och verkliga data påvisar att den förbättrade flexibiliteten hos kontextspecifika strukturer kan resultera i modeller med högre prediktiv förmåga än traditionella modeller.

List of original articles

I Pensar, J., Nyman, H., Koski, T. & Corander, J. (2015). Labeled directed acyclic graphs: a generalization of context-specific independence in directed graphical models. Data Mining and Knowledge Discovery **29**, 503–533.

II Pensar, J., Nyman, H., Lintusaari, J. & Corander, J. (2016). The role of local partial independence in learning of Bayesian networks. International Journal of Approximate Reasoning **69**, 91–105.

III Pensar, J., Nyman, H., Niiranen, J. & Corander, J. (2016). Marginal pseudo-likelihood learning of Markov network structures. Submitted.

IV Pensar, J., Nyman, H. & Corander, J. (2016). Structure learning of contextual Markov networks using marginal pseudo-likelihood. Submitted.

Authors' contributions to Articles I–IV

I The original idea is due to JC. All authors contributed to the development of the model class. JP and HN had the main responsibility in developing the score function. JP had the main responsibility in all remaining aspects of the article.

II JP had the main responsibility in all aspects of the article.

III JP had the main responsibility in all aspects of the article.

IV The original idea of applying the marginal pseudo-likelihood on contextual Markov networks is due to JP and JC. All authors contributed to the development of contextual Markov networks. JP had the main responsibility in all remaining aspects of the article.

In addition to the included articles, the author has co-authored the publications by Nyman et al. (2014, 2015a,b) and Janhunen et al. (2015), which are related to the work covered by this thesis.

Contents

1 Introduction

Probabilistic models provide a general tool for modeling real-world systems where there is a significant amount of uncertainty involved. In particular, in this thesis we consider (probabilistic) graphical models, which are used for compactly modeling complex joint distributions over a set of discrete variables. A compact representation of a potentially very high-dimensional distribution is achieved by exploiting structure in the distribution in the form of statements of conditional independence, which are naturally encoded by a graph structure. Characterized by the type of graph, the two most common families of graphical models are Bayesian networks and Markov networks, which are both considered in this thesis.

Graphical models have received considerable attention by the statistics and computer science community during the last few decades (Cowell et al., 1999; Koller & Friedman, 2009; Koski & Noble, 2009; Lauritzen, 1996; Pearl, 1988; Whittaker, 1990, among others). As a result of their generic applicability, graphical models have been applied in various fields and applications such as medical diagnosis, computer vision, analysis of genetic data, speech recognition, credit risk evaluation, computer security, and protein contact prediction.

Despite their wide adoption, the conditional-independence-based restrictions associated with traditional graphical models have been recognized to be unnecessarily coarse in certain situations. This observation has led to the development of more flexible models (Boutilier et al., 1996; Chickering et al., 1997; Corander, 2003; Friedman & Goldszmidt, 1996; Geiger & Heckerman, 1996; Højsgaard, 2003; Poole & Zhang, 2003). In particular, Boutilier et al. (1996) formalized the notion of context-specific independence (CSI) as a natural generalization of conditional independence. By including CSI into the graphical model framework, it is possible to obtain more accurate model structures which still enjoy a sound independence-based interpretation.

One of the main challenges related to graphical models is learning the model structure from data. This task is very demanding for several reasons, for example, the number of possible structures is extremely large. Still, from a user-perspective it is an important problem since the mere existence of complex models is of limited practical use, if they cannot be automatically and reliably learned from data. For this reason, there has been much research related to learning of graphical models (for an overview, see Koller & Friedman, 2009).

The main goals of the four articles included in this thesis are to generalize the concept of CSI in graphical models, develop efficient structure learning methods inspired by Bayesian statistics, and study learning of the proposed model classes in numerical experiments on both synthetic and real-world data. The introductory part of the thesis gives a brief overview of the work covered by the included articles in the context of related research. We begin in Section 2 by introducing the concept of graphical models and discussing the fundamental properties of both Bayesian networks and Markov networks. In Section 3, we introduce the notion of CSI and show how it can be included in the considered model classes as part of the model structures. In Section 4, we consider the structure learning problem by deriving score functions based on results from Bayesian statistics. In Section 5, we provide summaries of the included articles and discuss their contributions to the research field. Finally, in Section 6, we provide some concluding remarks and discuss potential future research.

2 Graphical models

We consider a set of d discrete random variables $X = \{X_1, \ldots, X_d\}$. Each variable X_j takes on values from a finite set of outcomes represented by $\mathcal{X}_j = \{0, 1, \ldots, r_j - 1\}$. We let $V = \{1, \ldots, d\}$ denote the indices of the variables. For a subset $S \subseteq V$, we denote the corresponding variables by X_S. We use $p(X)$ to denote the distribution over X, whereas $p(x)$ is shorthand for the probability $p(X = x)$.

The purpose of graphical models is to represent a joint distribution over X in an efficient and compact manner. Even in the case of binary variables, a naive representation requires $2^d - 1$ free parameters to specify a joint distribution over d variables. It is easy to realize that such a representation quickly becomes impractical as the number of variables is increased. To overcome this problem, graphical models break down the joint distribution into smaller more manageable parts by exploiting statements of *conditional independence*.

Definition 1. *Conditional Independence*
Let A, B, S be three disjoint subsets of V. We say that X_A is conditionally independent of X_B given X_S if

$$p(x_A \mid x_B, x_S) = p(x_A \mid x_S)$$

holds for all $(x_A, x_B, x_S) \in \mathcal{X}_A \times \mathcal{X}_B \times \mathcal{X}_S$ whenever $p(x_B, x_S) > 0$. This is denoted by

$$X_A \perp X_B \mid X_S.$$

If $S = \varnothing$, then $X_A \perp X_B$ is reduced to marginal independence between the two sets of variables.

To illustrate how conditional independence can be used in practice, consider the joint distribution over three binary variables $X = \{X_1, X_2, X_3\}$. Using the chain rule, the distribution can be factorized according to

$$p(X_1, X_2, X_3) = p(X_1)p(X_2 \mid X_1)p(X_3 \mid X_1, X_2). \tag{1}$$

Considering each factor individually, we need $1 + 2 + 4 = 7$ free parameters to specify the joint distribution. Now, assume that

$$X_2 \perp X_3 \mid X_1. \tag{2}$$

As stated in Definition 1, the last factor in (1) can be simplified accordingly such that

$$p(X_1, X_2, X_3) = p(X_1)p(X_2 \mid X_1)p(X_3 \mid X_1),$$

where X_2 no longer affects the conditional distribution of X_3 given X_1. The required number of free parameters is now reduced to $1 + 2 + 2 = 5$.

In the example above, the computational savings might seem negligible, however, when considering tens or even hundreds of variables, it would no longer be practically possible to model the joint distribution without simplifying assumptions. In such situations, it is no longer practical to represent the dependence structure among the variables in the form of a list of independence statements. Instead, the dependence structure of a graphical model is represented by a graph structure. The graph consists of *nodes* (or *vertices*) representing variables and *edges* representing direct dependences among the variables. On the other hand, lack of edges represents statements of conditional independence. The graph offers an intuitive way of illustrating the dependence structure to a human user. Moreover, it also enables use of graph

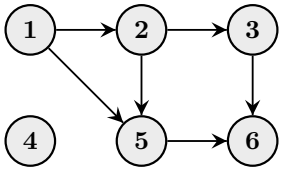


Figure 1: A DAG over six nodes.

theory when designing algorithms for learning and performing inference in graphical models.

There are two main families of graphical models; *Bayesian networks* (directed graphical models) and *Markov networks* (undirected graphical models). In this thesis, both types are considered. More specifically, Articles I–II consider Bayesian networks, while Articles III–IV consider Markov networks. A brief overview of the basic properties of each model class is given next. For a more detailed review of the theory of graphical models, see for example Koller & Friedman (2009).

2.1 Bayesian networks

The basis of the Bayesian network formulation is a *directed acyclic graph (DAG)*. We denote a DAG by $G = (V, E)$, where $V = \{1, \ldots, d\}$ is a set of nodes corresponding to the variables and E is a set of directed edges between the nodes such that (i, j) denotes a directed edge from node i to node j. The edge set must satisfy the *acyclicity* property, which means that starting from a node it is not possible to return to that node by following the direction of the edges. The *parents* of a node j are all nodes from which there is a directed edge to j, that is, $pa(j) = \{i \in V : (i, j) \in E\}$. A node j is called a *descendant* of node i, if it can be reached from node i following the direction of the edges. A *trail* is a sequence of nodes for which each pair of consecutive nodes are connected by an edge. As is typical in the graphical model literature, the terms node and variable are occasionally used interchangeably. An example of a DAG over six nodes is found in Figure 1.

In addition to the graph component, a Bayesian network specifies a joint distribution over the variables. The distribution must satisfy the conditional independence assumptions encoded by the DAG. These assumptions can be compactly characterized by the directed local Markov property. It states that each variable is conditionally independent of its non-descendants given its parents. Consequently, a DAG implies a factorization of the joint distribution according to

$$p(X_1, \ldots, X_d) = \prod_{j=1}^{d} p(X_j \mid X_{pa(j)}),$$

which is known as the *chain rule for Bayesian networks* (Koller & Friedman, 2009, p. 62). For example, the factorization according the DAG in Figure 1 is

$$p(X_1, \ldots, X_6) = p(X_1)p(X_2 \mid X_1)p(X_3 \mid X_2)p(X_4)p(X_5 \mid X_{1,2})p(X_6 \mid X_{3,5}).$$

The joint distribution of a Bayesian network is thus broken down over the nodes into local *conditional probability distributions (CPDs)*. Consequently, the probability of a joint configuration is simply determined by a product of factors, where each factor corresponds to a conditional probability of a variable given its parents. The basic,

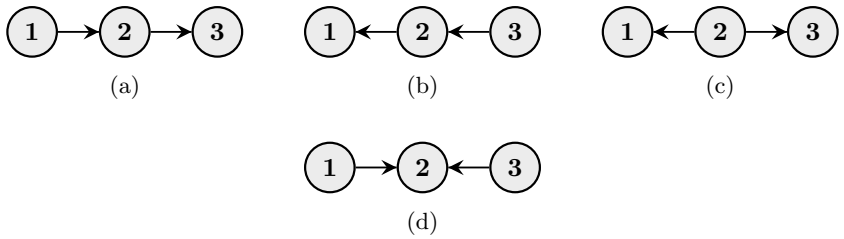


Figure 2: Possible connections in a Bayesian network: (a)–(b) *chain* connection, (c) *fork* connection, and (d) *collider* connection.

and perhaps most common, way of specifying the CPDs is in the form of *conditional probability tables (CPTs)*, which simply list the CPDs such that each row in the table represents a distinct parent configuration.

In addition to the local independences, a Bayesian network encodes a collection of non-local independences which can be derived from the local independences, however, such a derivation can be very cumbersome. Instead, non-local conditional independences can be verified directly from the graph by a sound procedure known as *d-separation*. When using d-separation, probabilistic influence should be considered as information flowing through the graph.

To illustrate d-separation and the fundamental properties of Bayesian networks, we look at the possible ways two nodes can be indirectly connected via a third node. The four possible connections are illustrated in Figure 2. Connections 2(a)–(c) are equivalent in the sense that information can pass between nodes 1 and 3 through node 2 if X_2 is not observed, while the flow is blocked by node 2 if X_2 is observed. This corresponds to nodes 1 and 3 being d-separated by node 2, which implies that the conditional independence statement

$$X_1 \perp X_3 \mid X_2$$

holds for graphs 2(a)–(c). In contrast, the collider connection in Figure 2(d) works in the opposite manner, that is, information can pass through node 2 only if X_2 is observed, while the flow is blocked by node 2 if X_2 is not observed. This corresponds to nodes 1 and 3 being d-separated by the empty set, which implies that the marginal independence statement

$$X_1 \perp X_3$$

holds for this graph, however, since nodes 1 and 3 are not d-separated by node 2, the former conditional independence is not implied by the graph. This is known as conditional dependence. The same reasoning as above carries over to more complicated graphs where there are more than one trail between a pair of nodes. More formally, a trail is referred to as *active* given S if all fork and chain nodes in the trail do not belong to S and for each collider node in the trail, either the collider node itself or one of its descendants belongs to S. Then, two nodes i and j are d-separated by S, implying that

$$X_i \perp X_j \mid X_S,$$

if there is no active trail between i and j given S.

The collider connection is also known as a *v-structure* and is a fundamental feature that separates Bayesian networks from undirected models. To further explain its behavior, we use a classic example from the Bayesian network literature (Pearl, 1988). Consider the graph in Figure 2(d). Let X_2 represent a newly installed burglar alarm,

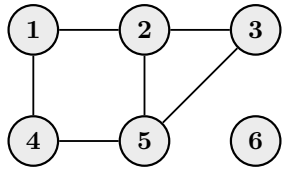


Figure 3: An undirected non-chordal graph over six nodes.

which reliably detects burglary, represented by X_1. However, the alarm tends to go off also in case of an earthquake, represented by X_3. There are now two possible and marginally independent causes that increase the probability of the alarm going off. Say that we are given information that the alarm has gone off, however, we also hear on the radio that there has been an earthquake in the area of the house. Since the knowledge of the earthquake in a sense explains the alarm going off, the alternative cause (burglary) is rendered less likely to have happened. This phenomenon is known as *explaining away*.

Since graphs 2(a)–(c) encode the same dependence structure, they are said to belong to the same *Markov equivalence class*. In terms of modeling a distribution, they are all equivalent in the sense that they represent the same set of distributions.

2.2 Markov networks

The dependence structure of a Markov network is represented by an undirected graph. We use the same graph notation $G = (V, E)$ as in the previous section with the difference that an undirected edge between node i and j is denoted by $\{i, j\}$. A *clique* C in a graph is defined as a subset of nodes for which all pairs of nodes are connected by an edge. A clique is defined as *maximal* if no additional node can be added to the clique without violating the clique criterion. The *Markov blanket* of a node j is denoted by $mb(j)$ and defined as all nodes which are connected to j by an edge. A *cycle* is a sequence of nodes that starts and ends with the same node and each pair of consecutive nodes are connected by an edge. A graph is said to be *chordal* if all cycles that contain four or more nodes have a *chord*, that is, an edge that is not part of the cycle but connects two nodes in the cycle. An example of an undirected graph over six nodes is found in Figure 3. This particular graph is non-chordal due to the chordless cycle $1 - 2 - 5 - 4 - 1$.

Similar to a Bayesian network, a Markov network specifies a joint distribution through a set of parameters associated with the undirected graph. However, in contrast to a Bayesian network, the parameters do not in general correspond to conditional probabilities or even probabilities, making them less intuitive. A common way of representing the joint distribution of a Markov network is through a factorization over the maximal cliques in the graph in terms of clique factors (see equation (2.1) of Article III). Another approach is to assume a positive distribution and represent the distribution in terms of a log-linear model

$$\log p(x_1, \ldots, x_d) = \sum_{A \subseteq V} \phi_A(x),$$

where the ϕ-terms are real-valued coordinate projection functions, such that $\phi_A(x) = \phi_A(x_A)$. In a graphical log-linear model, the ϕ-terms satisfy the graph related constraint

$$\phi_A(\cdot) = 0 \text{ if } \{i, j\} \subseteq A \text{ for some } \{i, j\} \notin E. \tag{3}$$

Furthermore, in order to avoid an over-parameterization, the ϕ-terms are also defined such that

$$\phi_A(x_A) = 0 \text{ if } x_j = 0 \text{ for any } j \in A. \tag{4}$$

As an example, the log-linear parameterization associated with the graph in Figure 3 is

$$\begin{aligned}\log p(x_1, \ldots, x_6) = \phi_\varnothing &+ \phi_1(x) + \phi_2(x) + \phi_3(x) + \phi_4(x) + \phi_5(x) + \phi_6(x) \\ &+ \phi_{1,2}(x) + \phi_{1,4}(x) + \phi_{2,3}(x) + \phi_{2,5}(x) + \phi_{3,5}(x) + \phi_{4,5}(x) \\ &+ \phi_{2,3,5}(x).\end{aligned}$$

As specified by restriction (3), no ϕ-term covers pair of nodes that are not in the edge set of the graph. For more details regarding the log-linear parameterization, see Whittaker (1990).

Similar to Bayesian networks, the graph of a Markov network encodes a dependence structure in terms of statements of conditional independence. The dependence structure can be characterized by the following Markov properties:

1. Pairwise Markov property: $X_i \perp X_j \mid X_{V\setminus\{i,j\}}$ for all $\{i, j\} \notin E$.
2. Local Markov property: $X_i \perp X_{V\setminus\{mb(i)\cup i\}} \mid X_{mb(i)}$ for all $i \in V$.
3. Global Markov property: $X_A \perp X_B \mid X_S$ for all disjoint subsets A, B, S of V such that S separates A from B.

The above properties are proven to be equivalent under the assumption of positivity of the joint distribution (Lauritzen, 1996). Note that the global Markov property is the undirected analogue of the d-separation criterion, however, here A and B are separated by S, if all paths between A and B pass through S.

2.3 Bayesian networks vs. Markov networks

Bayesian networks and Markov networks are in many respects similar. The end goal of both is to represent a joint distribution through a factorization justified by statements of conditional independence which are encoded by a graph. Conditional independences in either model can be verified directly from the graph through the use of separation criteria. The dependence structure encoded by an undirected graph is easier to interpret and perhaps more intuitive. On the other hand, the parameterization of a Bayesian network is more advantageous due to its "true" factorization. In comparison, the parameters of a (non-chordal) Markov network are connected through a normalizing constant known as the partition function, which corresponds to the $\phi_\varnothing$-term in the log-linear parameterization. The fundamental difference between Bayesian networks and Markov networks is that they can encode different dependence structures; Bayesian networks can represent conditional dependences through v-structures, whereas Markov networks can represent cyclic dependences. Still, the two model classes partially overlap since, for each chordal undirected graph, there is a corresponding class of Markov equivalent DAGs encoding the same dependence structure. In the end, each model class has its own strengths and weaknesses, and hence, which class is better suited for modeling a particular problem depends on the application in question.

3 Context-specific independence in graphical models

It has been noticed by several authors that conditional independence alone can in some situations be unnecessarily stringent for modeling real-world phenomena. In an attempt to loosen the restrictions associated with traditional graphical models, the notion of *context-specific independence (CSI)* has emerged (Boutilier et al., 1996; Corander, 2003; Friedman & Goldszmidt, 1996; Højsgaard, 2003; Poole & Zhang, 2003). CSI is a natural generalization of conditional independence and it was formalized by Boutilier et al. (1996) for the purpose of capturing regularities in the CPTs of Bayesian networks.

Definition 2. *Context-Specific Independence*
Let A, B, C, S be four disjoint subsets of V. We say that X_A is contextually independent of X_B given X_S and the context $X_C = x_C$ if

$$p(x_A \mid x_B, x_C, x_S) = p(x_A \mid x_C, x_S)$$

holds for all $(x_A, x_B, x_S) \in \mathcal{X}_A \times \mathcal{X}_B \times \mathcal{X}_S$ whenever $p(x_B, x_C, x_S) > 0$. This will be denoted by

$$X_A \perp X_B \mid x_C, X_S.$$

When comparing Definition 1 and 2, it is easy to realize that CSI is a more specific form of conditional independence in the sense that it only holds in part of the outcome space of the conditioning variables. In particular, we have the equivalence

$$X_A \perp X_B \mid x_C, X_S \text{ for all } x_C \in \mathcal{X}_C \Leftrightarrow X_A \perp X_B \mid X_C, X_S.$$

To illustrate the concept of CSI in practice, we return to our simple example in Section 2 concerning the factorization (1) of the joint distribution over three binary variables. However, instead of the conditional independence assumption (2), assume that

$$X_2 \perp X_3 \mid X_1 = 1 \text{ and } X_2 \not\perp X_3 \mid X_1 = 0.$$

This gives us a factorization of the joint distribution according to

$$p(X_1, X_2, X_3) = \begin{cases} p(X_1)p(X_2 \mid X_1)p(X_3 \mid X_1, X_2), & \text{if } X_1 = 0 \\ p(X_1)p(X_2 \mid X_1)p(X_3 \mid X_1), & \text{if } X_1 = 1 \end{cases}, \tag{5}$$

where X_2 no longer affects the conditional distribution of X_3 given the context $X_1 = 1$. The number of free parameters required to specify the distribution is reduced by one, from $1 + 2 + 4 = 7$ to $1 + 2 + 3 = 6$. If we were to only consider conditional independence, the above CSI statement could not be accounted for in the factorization. Consequently, by exploiting statements of CSI, it is possible to more accurately model a joint distribution without inducing redundant parameters.

One of the main goals of this thesis is to generalize the concept of CSI in graphical models, including both Bayesian networks (Article I) and Markov networks (Article IV). A brief overview of the introduced model classes and the theory surrounding them is given next. For more details, the reader is referred to the articles and other works referenced in the text.

(a) (b)

Figure 4: (a) A DAG and (b) a CSI-tree of node 3, which together represent the dependence structure corresponding to the factorization in (5).

3.1 Bayesian networks

The conditional independence assumptions made by a Bayesian network enable modeling high-dimensional joint distributions through node-wise CPDs. Still, the number of parameters required to specify a traditional CPT of a node j,

$$(|\mathcal{X}_j| - 1) \cdot |\mathcal{X}_{pa(j)}| = (r_j - 1) \cdot \prod_{i \in pa(j)} r_i,$$

may in some scenarios become overwhelming, since it grows exponentially with the number of parents. In order to counteract the rapid growth in the number of parameters, use of local CSI statements has been proposed and investigated by numerous authors (Boutilier et al., 1996; Friedman & Goldszmidt, 1996; Poole & Zhang, 2003). By local we refer to a statement concerning the relation between a node and its parents in the form of

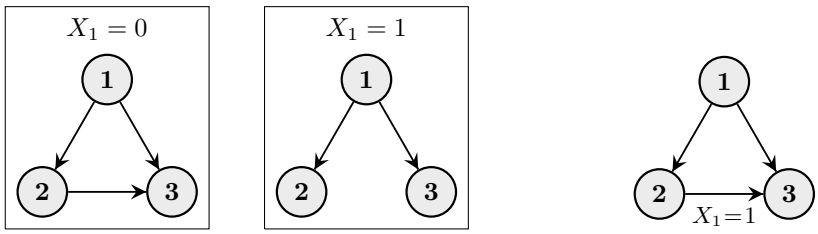
$$X_j \perp X_{pa(j)\setminus C} \mid x_C \text{ where } C \subset pa(j) \text{ (Definition 3, Article I).} \tag{6}$$

Local CSI statements are particularly well-suited for the Bayesian network parameterization since they imply that certain corresponding CPDs are identical. More specifically, they imply that

$$p(X_j \mid x_{pa(j)\setminus C}, x_C) = p(X_j \mid x'_{pa(j)\setminus C}, x_C)$$

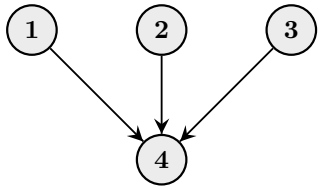
for all $x_{pa(j)\setminus C}, x'_{pa(j)\setminus C} \in \mathcal{X}_{pa(j)\setminus C}$. Since identical CPDs need only be specified once, the number of necessary model parameters can be reduced accordingly. This can be viewed as partitioning the outcome space of the parents into classes of configurations such that the conditional distribution of the node is invariant for parent configurations belonging to the same class.

To illustrate and capture local CSI statements, Boutilier et al. (1996) proposed using decision trees, here referred to as *CSI-trees*. The internal nodes in a CSI-tree are made up of the parents of the considered node, and the leaves represent distinct CPDs. By starting from the root and traversing down the tree until reaching a leaf, one obtains a context specifying a class in the partition of the parent outcome space. Any parents not included in the context are rendered contextually independent of the node given the context. For example, the complete DAG in Figure 4(a) and the CSI-tree over node 3 in Figure 4(b) together represent the factorization in (5). This representation thus requires a collection of CSI-trees in combination with a DAG. In Article II, we investigate the use of CSI-trees and similar graph structures. An alter-



(a) (b)

Figure 5: (a) A Bayesian multinet and (b) an LDAG, which both represent the dependence structure corresponding to the factorization in (5).



X_1	X_2	X_3	$p(X_4 \mid X_{1,2,3})$
0	0	0	p_1
0	0	1	p_1
0	1	0	p_2
0	1	1	p_3
1	0	0	p_4
1	0	1	p_4
1	1	0	p_4
1	1	1	p_4

(a) (b)

Figure 6: (a) A DAG over four nodes and (b) an example of a CPT of node 4.

native representation was proposed by Geiger & Heckerman (1996), who introduced the concept of *Bayesian multinets* which can represent asymmetric independence, such as CSI, by using multiple networks. For example, the dependence structure in Figure 4 can be represented by the two context-specific DAGs in Figure 5(a).

Inspired by the work of Corander (2003), in Article I we introduce *labeled directed acyclic graphs (LDAGs)* which specify the dependence structure using a single graphical structure. In an LDAG, each edge is assigned a label that specifies a set of contexts for which the influence of that edge "vanishes" according to local CSI statements. For example, the LDAG in Figure 5(b) captures the information from the graphs in 5(a) in a single structure. For clarity, the variable specifying the label is here explicitly specified. Given a fixed ordering of the nodes, this is not necessary since the variables specifying a label are all parents except the one that is part of the edge.

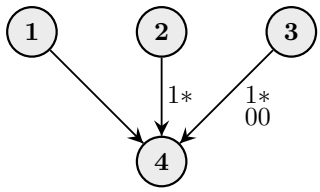
To further illustrate the concept of LDAGs, consider the DAG in Figure 6(a) and the associated CPT of node 4 in Figure 6(b). We assume that all variables are binary. A closer examination of the CPT reveals several identical CPDs which can be explained by CSI. Firstly, we see that $(0, 0, 0)$ and $(0, 0, 1)$ induce the same conditional distribution. This corresponds to the local CSI statement

$$X_4 \perp X_3 \mid X_1 = 0, X_2 = 0.$$

Furthermore, we see that the conditional distribution remains the same in the context $X_1 = 1$ regardless of the values of X_2 and X_3. This corresponds to

$$X_4 \perp \{X_2, X_3\} \mid X_1 = 1.$$

The above CSI statements can be turned into labels and all regularities in the CPT



(a)

X_1	X_2	X_3	$p(X_4 \mid X_{1,2,3})$
0	0	*	p_1
0	1	0	p_2
0	1	1	p_3
1	*	*	p_4

(b)

Figure 7: (a) An LDAG capturing the regularities in the CPT in Figure 6(b) and (b) the corresponding reduced CPT.

can be taken into account by the LDAG in Figure 7(a). Here we use a more compact notation where the label $1*$ represents the set $\{1\} \times \{0,1\} = \{(1,0),(1,1)\}$. Figure 7(b) contains the corresponding reduced CPT which has been constructed according to the labels such that each row represents a class in the partition of the parent outcome space:

$$
\begin{aligned}
&0\ 0\ * = \{(0,0,0),(0,0,1)\},\\
&0\ 1\ 0 = \{(0,1,0)\},\\
&0\ 1\ 1 = \{(0,1,1)\},\\
&1\ *\ * = \{(1,0,0),(1,0,1),(1,1,0),(1,1,1)\}.
\end{aligned}
$$

For more details on how the reduced CPT (or partition) is constructed with respect to an LDAG, see Section 2.1 of Article I.

The regularities in the above CPT are again consistent with a CSI-tree, however, this is not always the case. In contrast, LDAGs are general representations of CSI in the sense that they can represent any collection of CPD regularities consistent with CSI. Consequently, any Bayesian network with CSI-trees can be represented by an LDAG, whereas all LDAGs cannot be represented using CSI-trees. For an example of this, see Figure 5 of Article I.

In Section 2.2 of Article I, the properties of LDAGs are investigated more in detail and corresponding context-specific versions of d-separation and Markov equivalence are introduced and discussed. As an example of different LDAGs representing the same dependence structure, change the direction of the edge between nodes 2 and 3 in Figure 5(b).

3.2 Markov networks

Although originally formalized in the context of Bayesian networks, the notion of CSI has also been investigated as a means for improving the flexibility of Markov networks (Corander, 2003; Højsgaard, 2003; Nyman et al., 2014, 2015a,b). In particular, Corander (2003) introduced the class of *labeled graphical models*, which later was investigated further by Nyman et al. (2014, 2015a) who developed the subclasses of *decomposable stratified graphical models* and *stratified graphical models*. Compared to the original work by Corander (2003), certain restrictions were imposed on the stratified graphical model classes in order to facilitate the model learning process. In Article IV, we introduce the class of *contextual Markov networks* which in principle is equivalent to the class of labeled graphical models, however, it is defined in a slightly different manner due to an observation made in Nyman et al. (2015a).

Similar to LDAGs, each edge in a contextual Markov network is assigned a context (or label) which is now specified by the *common neighbors* of the edge nodes. The

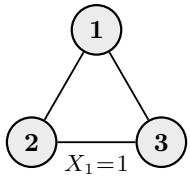


Figure 8: Labeled undirected graph encoding the dependence structure of a contextual Markov network.

common neighbors with respect to an undirected edge $\{i, j\}$ are denoted and defined by $cn(i, j) = mb(i) \cap mb(j)$. An edge context specifies values for which the direct dependence encoded by the edge "vanishes" according to local CSI statements which are now of the form

$$X_i \perp X_j \mid x_{cn(i,j)}, X_{V \setminus \{cn(i,j) \cup \{i,j\}\}} \text{ (Definition 2, Article IV).} \tag{7}$$

Using the common neighbors to specify an edge context is proven to be a natural condition. In Section 2.2 of Article IV, we show that the generality of contextual Markov networks would not be increased if allowing an edge context to be specified by supersets or subsets of the common neighbors.

The edge contexts can be illustrated by labeled undirected graphs using a similar notation as for LDAGs. Following a similar reasoning as Corander (2003) and Nyman et al. (2015a), we show in Proposition 2 of Article IV that CSI statements of the above type correspond to linear restrictions among the log-linear parameters. To illustrate the idea behind the result and at the same time bridge the gap between CSI in Bayesian networks and CSI in Markov networks, we return to our toy network. Consider the labeled undirected graph in Figure 8. Again for clarity, we have explicitly stated that X_1 specifies the label although it is clear since node 1 is the common neighbor of nodes 2 and 3, or using our notation, $cn(2, 3) = 1$. The above labeled graph obviously encodes the same dependence structure as the LDAG in Figure 5(b). The underlying complete graphs are equivalent and, according to (6) and (7), both labels encode the CSI statement

$$X_2 \perp X_3 \mid X_1 = 1.$$

The question is then: How is the above CSI restriction taken into account in the log-linear parameterization? From the LDAG framework, we know that the CSI statement implies that

$$p(X_3 \mid X_1 = 1, X_2 = 0) = p(X_3 \mid X_1 = 1, X_2 = 1).$$

However, the above restriction can after some reformulation be expressed in terms of joint probabilities:

$$\begin{aligned} \frac{p(X_3 = 0 \mid X_1 = 1, X_2 = 0)}{p(X_3 = 1 \mid X_1 = 1, X_2 = 0)} &= \frac{p(X_3 = 0 \mid X_1 = 1, X_2 = 1)}{p(X_3 = 1 \mid X_1 = 1, X_2 = 1)} \\ &\Leftrightarrow \\ \frac{p(X_3 = 0, X_1 = 1, X_2 = 0)}{p(X_3 = 1, X_1 = 1, X_2 = 0)} &= \frac{p(X_3 = 0, X_1 = 1, X_2 = 1)}{p(X_3 = 1, X_1 = 1, X_2 = 1)} \end{aligned}$$

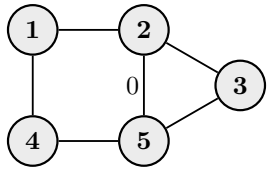


Figure 9: An undirected non-chordal labeled graph over five nodes.

Taking the logarithm of both sides and using the log-linear expansion, we obtain

$$(\phi_\varnothing + \phi_1) - (\phi_\varnothing + \phi_1 + \phi_3 + \phi_{1,3})$$
$$=$$
$$(\phi_\varnothing + \phi_1 + \phi_2 + \phi_{1,2}) - (\phi_\varnothing + \phi_1 + \phi_2 + \phi_3 + \phi_{1,2} + \phi_{1,3} + \phi_{2,3} + \phi_{1,2,3}).$$

We drop the arguments from the ϕ-terms since we are dealing with binary variables under assumption (4). After simplification the above equation is reduced to

$$\phi_{2,3} + \phi_{1,2,3} = 0,$$

which elegantly captures the considered CSI. The same reasoning was used in a more general setting to prove Proposition 2 of Article IV which states that a contextual Markov network can be formulated in terms of a log-linear model in which the parameters are being subject to linear restrictions implied by the edge contexts.

As mentioned, the class of contextual Markov networks is basically equivalent to the class of labeled graphical models, except that the definition of edge context (or label) has been modified to remain sound for non-chordal graphs. In previous works (Corander, 2003; Nyman et al., 2014, 2015a), a label was defined to encode CSI statements of the form

$$X_i \perp X_j \mid x_{cn(i,j)}. \tag{8}$$

In a chordal graph, the common neighbors are indeed sufficient to cut off any indirect dependences, however, consider the non-chordal graph in Figure 9. Even if the direct dependence between node 2 and 5 is removed and the indirect dependence via $2-3-5$ is blocked by node $cn(2,5) = 3$, there is still a path $2-1-4-5$ along which information can flow, rendering the variables dependent. To address this issue in Article IV, the new definition (7) explicitly refers to the direct dependence by conditioning on the remaining network (cf. pairwise Markov property).

4 Structure learning of graphical models

In this section, we consider one of the main inference tasks related to graphical models, learning the model structure from a set of data. The task of constructing networks manually is at best daunting and in general infeasible due to the complexity of the models. It is therefore of utmost importance to develop efficient learning algorithms that can automatically identify a suitable model from data. From a practical standpoint, developing more expressive model classes is of limited use if the models cannot be learned from data. With this in mind, in addition to developing new and more flexible model classes, the main emphasis of this thesis is on learning the structure of such models.

As will be discussed next, Bayesian networks and Markov networks pose different problems in the learning phase. Moreover, the learning task is already challenging for traditional graphical models and generalizing the models in terms of CSI further complicates the matter. Still, the potential gain of a more flexible model is that it can better emulate a target distribution without inducing redundant parameters.

4.1 Score-based learning

Structure learning methods can roughly be divided into two categories; constraint-based and score-based. Constraint-based methods try to infer the dependence structure through a series of separate independence tests that exploit the fundamental independence assumptions associated with the model class. Score-based methods, on the other hand, approach the learning task as an optimization problem over the space of possible model structures. Firstly, this requires a score function by which the plausibility of each structure can be evaluated. Secondly, to find high-scoring networks, this also requires an optimization algorithm since an exhaustive evaluation of the structure space is in general infeasible beyond toy-sized systems. Score-based methods are usually more demanding computationally, however, they tend to be more stable since they adopt a more global approach.

In this thesis, we focus on scored-based methods. The score functions are derived according to a Bayesian view and optimized using both a stochastic algorithm (Article I) and various deterministic algorithms (Articles II–IV). In the coming sections, we focus on the derivation of the score functions and mainly discuss the optimization in terms of how the structure of the scores can be exploited to design efficient search algorithms. For more details on the specific search algorithms, the reader is referred to the included articles.

4.2 Dirichlet as conjugate for the categorical distribution

Before proceeding to discuss the learning methods, we will go through a standard result from Bayesian analysis which is central for the derivation of the Bayesian score functions used in this thesis. The result concerns a special relationship between the categorical and Dirichlet distributions.

Definition 3. *Categorical distribution*
A categorical distribution over a discrete variable X with $r > 0$ possible outcomes and parameters $\theta = (\theta_1, \ldots, \theta_r)$, where $\theta_1, \ldots, \theta_r > 0$ and $\sum_{i=1}^{r} \theta_i = 1$, has the probability mass function $p(X = x^{(i)}; \theta) = \theta_i$ for $i = 1, \ldots, r$.

The categorical distribution is a generalization of the Bernoulli distribution ($r = 2$) and is the most general distribution over an r-way outcome space since the probability of each outcome is separately defined through $\theta = (\theta_1, \ldots, \theta_r)$. If denoting the

outcome by a vector rather than an integer, the categorical distribution is equivalent to a multinomial distribution over a single trial. As a result of this, the categorical distribution is often also referred to as the multinomial distribution in the literature.

Definition 4. *Dirichlet distribution*
A Dirichlet distribution over $r \geq 2$ variables $\theta = (\theta_1, \ldots, \theta_r)$ with parameters $\alpha = (\alpha_1, \ldots, \alpha_r)$, where $\alpha_i > 0$ for $i = 1, \ldots, r$, has a probability density function given by

$$f(\theta; \alpha) = \frac{\Gamma(\sum_{i=1}^{r} \alpha_i)}{\prod_{i=1}^{r} \Gamma(\alpha_i)} \prod_{i=1}^{r} {\theta_i}^{\alpha_i - 1}$$

if $\theta_1, \ldots, \theta_r > 0$ and $\sum_{i=1}^{r} \theta_i = 1$. The density is zero elsewhere.

The Dirichlet distribution is a multivariate generalization of the beta distribution ($r = 2$). The support of the Dirichlet distribution is the $(r-1)$-dimensional simplex, which basically is a set of r-dimensional generic discrete probability distributions, that is, categorical distributions.

The reason for the popularity of the Dirichlet distribution in Bayesian statistics is that it is a *conjugate prior* for the categorical (and multinomial) distribution. A distribution is called a conjugate prior for the *likelihood function* if the *posterior* and *prior* distributions belong to the same family of distributions. More specifically, let $\mathbf{x}$ denote a sample of n i.i.d. observations assumed to have been generated from

$$X \mid \theta \sim \text{Categorical}(\theta)$$

and let n_i denote the number of times outcome i occurs in the dataset. The posterior distribution $\theta \mid \mathbf{x}$ is defined as

$$f(\theta \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \theta) f(\theta)}{p(\mathbf{x})}, \tag{9}$$

where $p(\mathbf{x} \mid \theta)$ is the likelihood function of the parameters for the given data, in our case given by

$$p(\mathbf{x} \mid \theta) = \prod_{i=1}^{r} \theta_i^{n_i},$$

$f(\theta)$ is the prior distribution over the parameters, and

$$p(\mathbf{x}) = \int p(\mathbf{x} \mid \theta) f(\theta) d\theta \tag{10}$$

is the probability of the data, known as the *marginal likelihood*. Now, assuming that the prior distribution over the parameters is Dirichlet,

$$\theta \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_r),$$

then the corresponding posterior distribution is also Dirichlet,

$$\theta \mid \mathbf{x} \sim \text{Dirichlet}(\alpha_1 + n_1, \ldots, \ldots, \alpha_r + n_r),$$

where the α's, known as *hyperparameters*, have been updated by the counts in the data. In this context, it is easy to realize why the hyperparameters in the Dirichlet prior are also commonly referred to as *pseudo-counts*.

From a computational perspective, a conjugate prior is very convenient since it results in a closed-form expression for the posterior. Moreover, it allows us to derive a closed-form expression for the marginal likelihood:

$$p(\mathbf{x}) = \frac{\Gamma(\sum_{i=1}^{r} \alpha_i)}{\Gamma(n + \sum_{i=1}^{r} \alpha_i)} \prod_{i=1}^{r} \frac{\Gamma(n_i + \alpha_i)}{\Gamma(\alpha_i)}. \tag{11}$$

The above formula is the key result used when deriving the Bayesian score functions at which we look at next. In practice, the logarithm of the above formula is used, since it is computationally more manageable.

4.3 Marginal likelihood for Bayesian networks

The most widely used score for structure learning of Bayesian networks from data is the Bayesian score. By a dataset $\mathbf{x}$, from now on we refer to a complete dataset consisting of n i.i.d. joint observations over d variables. In the Bayesian approach, a graph G is scored by the unnormalized conditional probability of the graph given a dataset $\mathbf{x}$,

$$p(G \mid \mathbf{x}) \propto p(\mathbf{x} \mid G)p(G),$$

where $p(\mathbf{x} \mid G)$ is the marginal likelihood under the given graph and $p(G)$ is the prior probability of the graph.

The key factor of the Bayesian score is the marginal likelihood which, similar to (10), is evaluated by

$$p(\mathbf{x} \mid G) = \int p(\mathbf{x} \mid G, \theta) f(\theta \mid G) d\theta. \tag{12}$$

We let $\theta_{jl} = (\theta_{1jl}, \ldots, \theta_{r_j jl})$ be the parameters specifying the CPD (in terms of a categorical distribution) over node j given that the parents have been assigned configuration l. Furthermore, let n_{ijl} be the count of the number of times the corresponding family configuration occurs in the data. The likelihood function can then be compactly represented by the product

$$p(\mathbf{x} \mid G, \theta) = \prod_{j=1}^{d} \prod_{l=1}^{q_j} \prod_{i=1}^{r_j} \theta_{ijl}^{n_{ijl}}. \tag{13}$$

Under certain assumptions listed by Heckerman et al. (1995), the above integral can be solved analytically resulting in a closed-form expression (Buntine, 1991; Cooper & Herskovitz, 1992). In particular, one of the key assumptions is parameter independence, which allows for a factorization of the parameter prior according to

$$p(\theta \mid G) = \prod_{j=1}^{d} \prod_{l=1}^{q_j} f(\theta_{jl}). \tag{14}$$

This means that the global integral in (12) can be replaced by a product of local integrals. By further assuming Dirichlet distributions over the θ_{jl}'s,

$$\theta_{jl} \sim \text{Dirichlet}(\alpha_{1jl}, \ldots, \alpha_{r_j jl}),$$

the local integrals can be solved in an equivalent manner as (11). The marginal likelihood for Bayesian networks is then obtained as the closed-form expression

$$p(\mathbf{x} \mid G) = \prod_{j=1}^{d} \prod_{l=1}^{q_j} \frac{\Gamma(\sum_{i=1}^{r_j} \alpha_{ijl})}{\Gamma(n_{jl} + \sum_{i=1}^{r_j} \alpha_{ijl})} \prod_{i=1}^{r_j} \frac{\Gamma(n_{ijl} + \alpha_{ijl})}{\Gamma(\alpha_{ijl})}, \tag{15}$$

(a) (b)

Figure 10: Two DAGs, (a) G_1 and (b) G_2, over six nodes that are identical except that the direction of the edge between node 2 and 3 is different.

where $n_{jl} = \sum_{i=1}^{r_j} n_{ijl}$. The above expression is easily evaluated once the hyperparameters have been specified.

In addition to the marginal likelihood, the Bayesian score includes a graph prior, $p(G)$, through which it is possible to incorporate prior beliefs regarding the graph in terms of, for example, degree of sparsity. To maintain the useful factorization of the final score, the prior must be defined in a way that enables such a factorization. In general, the marginal likelihood as such has empirically been shown to give good results for Bayesian networks with standard CPTs. Therefore, it is quite common to assume a uniform prior over the graph space.

Note that the marginal likelihood factorizes into node-wise marginal conditional likelihoods according to

$$p(\mathbf{x} \mid G) = \prod_{j=1}^{d} p(\mathbf{x}_j \mid \mathbf{x}_{pa(j)}). \tag{16}$$

For example, the marginal likelihood for the graph in Figure 10(a) is factorized as

$$p(\mathbf{x} \mid G_1) = p(\mathbf{x}_1)p(\mathbf{x}_2 \mid \mathbf{x}_1)p(\mathbf{x}_3 \mid \mathbf{x}_2)p(\mathbf{x}_4)p(\mathbf{x}_5 \mid \mathbf{x}_{1,2})p(\mathbf{x}_6 \mid \mathbf{x}_{3,5}).$$

The factorization property makes the marginal likelihood attractive from a learning perspective. In particular, search algorithms based on local edge changes (add, delete, and reverse edge) exploit this. In order to evaluate a single edge change, it suffices to re-evaluate at most two node-wise scores since the remaining are kept identical and can be re-used from the previous iteration. For example, say that we want to compare the two graphs in Figure 10 which are otherwise identical but the direction of the edge between node 2 and 3 is different. To compare the graphs (under a uniform prior), we calculate the ratio of their marginal likelihoods, known as the *Bayes factor*, which is reduced to

$$\frac{p(\mathbf{x} \mid G_1)}{p(\mathbf{x} \mid G_2)} = \frac{p(\mathbf{x}_2 \mid \mathbf{x}_1)p(\mathbf{x}_3 \mid \mathbf{x}_2)}{p(\mathbf{x}_2 \mid \mathbf{x}_{1,3})p(\mathbf{x}_3)}$$

since the remaining factors cancel out. Given that G_1 is our current graph and that we have stored the values of its node-wise factors, it is sufficient to calculate the two new factors in the denominator in order to evaluate the above expression. Add and delete operations are even simpler to evaluate, since they only affect the local structure of a single node. The search algorithms in Articles I–II are designed to exploit this factorization property.

Another attractive property of the marginal likelihood for Bayesian networks is that it can readily be modified to take CSI and similar local independences into account (Chickering et al., 1997; Friedman & Goldszmidt, 1996). In a similar manner as previous works, we modified the marginal likelihood to cover LDAGs (Article I)

and networks where the CPTs are modeled using various graph-based representations (Article II). The common thing for these generalized networks is that they partition each parent outcome space into classes with invariant CPDs for parent configurations contained by the same class. Consequently, the marginal likelihood can still be evaluated by expression (15), however, with the distinction that the l-index now runs over parent classes rather than distinct parent configurations.

While the marginal likelihood works well as such for traditional Bayesian networks, we noticed in Article I that it favored dense LDAGs with large label sets. The complex structures of such models are not only computationally demanding to learn but the models also showed tendencies of overfitting manifested in poor out-of-sample predictive performance. To attend this issue, a tunable prior that penalized inclusion of labels was proposed. The tuning parameter was chosen by a cross-validation-based method. In Article II, the observation in Article I was confirmed in more extensive simulation studies. In this article, we designed a prior that promoted sparsity in terms of the graph structure.

In Articles I–II, we show how to further exploit the structural decomposition of the marginal likelihood when learning parent classes through local changes in an analogous manner as discussed above. Since the marginal likelihood score for a node j is further factorized over the parent classes,

$$p(\mathbf{x}_j \mid \mathbf{x}_{pa(j)}) = \prod_{l=1}^{q_j} p(\mathbf{x}_j \mid \mathbf{x}_{pa(j)}^{(l)}),$$

it is sufficient to only re-evaluate those classes that have been modified and re-use the scores for the remaining classes from the previous iteration.

4.4 Marginal pseudo-likelihood for Markov networks

Whereas the marginal likelihood can be evaluated in closed form for Bayesian networks, its calculation poses significant problems for Markov networks. Due to the partition function, likelihood-based scores are in general intractable for non-chordal Markov networks. For this reason, alternative objective functions have been proposed. One of the most popular is perhaps the *pseudo-likelihood* introduced by Besag (1975). In Article III, we introduce the *marginal pseudo-likelihood (MPL)* as a Bayesian version of the pseudo-likelihood score.

The pseudo-likelihood approximates the likelihood by a product of conditional likelihoods over each individual node conditional on all other variables or, given a graph, the Markov blankets,

$$\hat{p}(\mathbf{x} \mid G, \theta) = \prod_{j=1}^{d} p(\mathbf{x}_j \mid \mathbf{x}_{mb(j)}, \theta).$$

In a similar manner as in the previous section, we use the notation θ_{ijl} to represent the conditional probability of variable j being assigned value i given that the node's Markov blanket $mb(j)$ has been assigned configuration l. We modify the definition of the count n_{ijl} analogously. Given the modified notation, the pseudo-likelihood of a graph G can be expressed as

$$\hat{p}(\mathbf{x} \mid G, \theta) = \prod_{j=1}^{d} \prod_{l=1}^{q_j} \prod_{i=1}^{r_j} \theta_{ijl}^{n_{ijl}},$$

Figure 11: (a) An undirected graph and (b) a labeled undirected graph over six nodes.

which has a striking resemblance to the likelihood in (13). The MPL is obtained by replacing the likelihood in (12) with the pseudo-likelihood. By assuming a similar factorization of the parameter prior as in (14), we can solve the MPL in closed form obtaining a similar expression as in (15), however, the l-index runs over Markov blanket configurations instead of parent configurations. See Section 4.1 of Article III for more details.

It is worth pointing out that the parameter independence assumption made during the derivation of the MPL is justified purely by computational convenience, since it actually violates the properties of a distribution associated with a Markov network. Still, in Theorem 4.1 of Article III, we motivate the MPL from a theoretical standpoint by establishing consistency in the large sample limit, that is, the correct graph will obtain the highest score as the sample size tends to infinity.

Similar to the marginal likelihood for Bayesian networks (16), the MPL factorizes into disconnected marginal conditional likelihoods,

$$\hat{p}(\mathbf{x} \mid G) = \prod_{j=1}^{d} p(\mathbf{x}_j \mid \mathbf{x}_{mb(j)}).$$

As an example, the undirected graph in Figure 11(a) is evaluated according to

$$\hat{p}(\mathbf{x} \mid G) = p(\mathbf{x}_1 \mid \mathbf{x}_{2,4})p(\mathbf{x}_2 \mid \mathbf{x}_{1,3,5})p(\mathbf{x}_3 \mid \mathbf{x}_{2,5})p(\mathbf{x}_4 \mid \mathbf{x}_{1,5})p(\mathbf{x}_5 \mid \mathbf{x}_{2,3,4})p(\mathbf{x}_6).$$

The factorization property makes the MPL an attractive objective function from a computational perspective. Search algorithms based on local changes (add/delete edge) are particularly convenient since a single edge change will modify the Markov blankets of only two nodes. Consequently, only two node-wise scores need to be re-evaluated whereas the remaining scores are kept unchanged and can be re-used from the previous iteration. For more details regarding this, see Sections 4.3 and 5.2 of Article III.

The main reason for imposing restrictions, such as chordality, on the context-specific model classes in Nyman et al. (2014, 2015a) is to facilitate the model learning process. For this purpose, we extended the scope of the MPL in Article IV to also cover contextual Markov networks. In particular, by combining the results of Articles I and III, we showed that the MPL can still be evaluated in closed form for this general class of context-specific Markov networks. The key innovation lies in the observation that the CSI statements of a contextual Markov network can be accounted for by the MPL by partitioning the outcome space of the Markov blankets in a similar manner as the outcome space of the parents in an LDAG. To give an example, consider the labeled graph in Figure 11(b). The label represents the CSI statement

$$X_2 \perp X_5 \mid X_3 = 1, X_{1,4,6}.$$

Under the Markov properties of the given graph, the CSI can be reformulated according to

$$X_2 \perp X_5 \mid X_3 = 1, X_{1,4,6} \Leftrightarrow \begin{matrix} X_2 \perp X_5 \mid X_3 = 1, X_1 \\ X_5 \perp X_2 \mid X_3 = 1, X_4 \end{matrix},$$

where the statements on the right are analogous to local CSI statements in an LDAG where a node's Markov blanket is thought of as parents. Consequently, the statements on the right can readily be accounted for by the MPL by partitioning the outcome space of the Markov blankets $mb(2) = \{1, 3, 5\}$ and $mb(5) = \{2, 3, 4\}$ accordingly when evaluating the scores of node 2 and node 5. The scores of the remaining nodes are calculated as regular MPL where each Markov blanket configuration is considered separately. For more details, see Section 3.2 of Article IV.

As for traditional Markov networks, we show in Theorem 1 of Article IV that the resulting MPL-based estimator is consistent in selecting the structure of a contextual Markov network. To avoid the issue of identifying overly dense graphs for limited sample sizes, we propose a tunable prior that penalizes inclusion of context elements (or labels) in a similar fashion as the prior proposed for LDAGs. However, in contrast to the cross-validation-based approach used in Article I, we use the *Bayesian information criterion (BIC)* (Schwarz, 1978), which is also a consistent score, for choosing the final model structure from a collection of candidate structures learned under differently tuned priors. The idea behind the approach is that any potential overfitting with respect to the pseudo-likelihood-based MPL will lead to a reduced value on the maximum-likelihood-based BIC score, the numerical experiments showed that the method seems to work quite well in practice. The obvious advantage of using BIC instead of cross-validation is that the maximum likelihood estimates of the log-linear model parameters only need to be calculated once, which is a computationally demanding task due to the partition function.

5 Summaries of the included articles

5.1 Article I: Labeled directed acyclic graphs: a generalization of context-specific independence in directed graphical models

This article introduces the concept of labeled directed acyclic graphs (LDAGs) as a tool for representing context-specific independence (CSI) in Bayesian networks. In contrast to previous proposals, we show that LDAGs can represent general CSI constraints through a single graph structure. We introduce and discuss several properties of LDAGs in terms of model identifiability and interpretability. To facilitate the interpretation of LDAGs, we re-use conditions originally introduced for the class of labeled graphical models (Corander, 2003). Based on the work by Boutilier et al. (1996), we introduce and discuss a context-specific version of the d-separation criterion which can be applied on LDAGs. Finally, we investigate situations where two distinct LDAGs can represent the same dependence structure.

To enable efficient learning of LDAGs from data, we derive a Bayesian score for which the marginal likelihood can be calculated analytically. This is achieved by assuming an LDAG-based factorization of the Dirichlet prior for the model parameters in a similar manner as Friedman & Goldszmidt (1996). To identify high-scoring structures, we use a stochastic search, developed in Corander et al. (2008, 2006), combined with a deterministic greedy hill-climb method. During the numerical simulations, we noticed that the marginal likelihood alone has a tendency of overfitting by favoring dense LDAGs with large label sets. For this reason, we propose a tunable structure prior which penalizes inclusion of labels. To choose among several candidate values on the tuning parameter, we use a cross-validation-based method.

In our simulations, we use synthetic Bayesian networks based on both a DAG and an LDAG. The quality of an identified structure is assessed by the Kullback-Leibler divergence between the true distribution and the approximate model distribution. The numerical experiments show that the models based on LDAGs can outperform traditional Bayesian networks in terms of approximating an actual network distribution, especially when the true network contains CSI.

5.2 Article II: The role of local partial independence in learning of Bayesian networks

This article further investigates the role of structured conditional probability tables (CPTs) when learning Bayesian networks. We consider models with various degrees of expressiveness, from restrictions consistent with CSI to arbitrary equalities among the conditional probability distributions. To collect all such restrictions under a single notion, we introduce the concept of partial conditional independence. Due to computational advantages, we focus on tree-like CPT structures. In particular, we show that CSI-trees can be extended to capture an additional form of regularities, which are particularly useful for high cardinality variables.

To evaluate the plausibility of the model structures, we modify the Bayesian score from Article I in a similar manner as Chickering et al. (1997). However, in contrast to the label-dependent prior in Article I, we define a structure prior in terms of the DAG alone and do not distinguish between different CPT structures *a priori*. To identify high-scoring models, we use a deterministic search algorithm which traverses greedily among DAGs using local edge changes. The CPT structures are learned using a greedy hill-climb method that operates in a top-down fashion.

We perform extensive numerical experiments on both synthetic data generated by benchmark Bayesian networks and real data from a machine learning repository. To assess the quality of the models, we use, among others, a measure of predictive accuracy which is comparable to empirical Kullback-Leibler divergence estimated from observed data. We show that including CPT structures in the learning process may significantly improve the quality of the inferred models for both synthetic and real data. However, we also confirm our observation from Article I in that it is usually necessary to further regulate the marginal likelihood through, for example, a prior over the network structures.

5.3 Article III: Marginal pseudo-likelihood learning of Markov network structures

This article introduces a new Bayesian-type score function for learning the graph structure of non-chordal Markov networks. Due to the partition function, the Bayesian approach for learning the graph structure from data has been restricted to chordal Markov networks for which the marginal likelihood can be calculated analytically (Dawid & Lauritzen, 1993). Chordality, however, is a rather strong assumption which may be unnatural when modeling real-world phenomena. Therefore, we introduce the marginal pseudo-likelihood (MPL) as a Bayesian version of the pseudo-likelihood (Besag, 1975) where graph-specific nuisance parameters are marginalized out.

We show that the MPL can be evaluated in closed form under certain assumptions. We investigate the properties of the MPL as a scoring function and, in particular, we show in Theorem 4.1 that the resulting MPL-based graph estimator is consistent in the large sample limit. We discuss the computational complexity of the MPL and its attractiveness from an optimization perspective. Finally, we also discuss the relationship between MPL and the asymptotically equivalent *pseudo-Bayesian information criterion* (Csiszár & Talata, 2006) and a special class of *dependency networks* (Heckerman et al., 2001).

For MPL optimization, we design a two-step procedure which can be applied on high-dimensional systems. The first step works as a pre-scan picking out potential edges and the second step performs a greedy hill-climb on a restricted graph space determined by the first step. We perform extensive experiments comparing our MPL method to several competing methods on both synthetic and real-world networks with known graph structure. The performance of the methods is evaluated by the resemblance between the inferred and true graph structure as quantified by the *Hamming distance*. Overall, the MPL method outperforms the competing methods at a comparable learning time.

5.4 Article IV: Structure learning of contextual Markov networks using marginal pseudo-likelihood

This article introduces a general class of context-specific Markov networks, called contextual Markov networks. Context-specific Markov networks were originally introduced by Corander (2003) and later further developed by Nyman et al. (2014, 2015a). One of the main challenges with these models has been the task of learning the model structure from data. For this reason, Nyman et al. (2014) introduced restrictions on the models in the form of chordality and certain context-related conditions, which together allow for the marginal likelihood to be evaluated in closed form. In Nyman et al. (2015a), the restrictions on the context structure were lifted making the models more flexible, but at the same time likelihood-based scores intractable in

practice for larger systems. In this article, we lift the restriction of chordality and consider a fully general setting in terms of CSI, as originally proposed by Corander (2003).

The main contribution of this article is extending the scope of MPL to contextual Markov networks by combining the results from Articles I and III. We show that the MPL can still be evaluated in closed form, since the considered CSI statements can be accounted for in a similar manner as local CSI statements in LDAGs. Furthermore, we show that the MPL-based estimator for contextual Markov network structures is consistent in the large sample limit. To avoid the issue of overfitting, we propose a similar tunable prior as was used in Article I, however, instead of using cross-validation, we choose the final model according to the Bayesian information criterion (Schwarz, 1978).

To identify high-scoring structures, we design a deterministic greedy hill-climb algorithm. We perform numerical experiments to investigate how the MPL performs in practice on both synthetic and real-world data. The identified structures are primarily evaluated by the predictive accuracy of the corresponding models. The model parameters are approximated by the maximum likelihood estimates which are calculated by a conjugate gradient ascent technique. Overall, the identified contextual Markov networks show an improved predictive accuracy both in- and out-of-sample compared to traditional Markov networks.

6 Concluding remarks and future research

The notion of CSI has been proposed as a means to generalize probabilistic graphical models such as Bayesian networks and Markov networks. We have further pursued this idea through the concept of context-specific graphical models in which CSI is included as part of the model structure. The main emphasis of this thesis has been on learning such model structures from data. Compared to traditional graphical models, learning the structure of context-specific graphical models is considerably more challenging due to the extremely large space of possible structures. In addition to making the learning task more demanding computationally, we noticed a previously not recognized problem in the form of overfitting if the structure was optimized with respect to the marginal likelihood alone. To fix this issue, we proposed using structure priors to further regulate the model fit.

In terms of learning Bayesian networks, the Bayesian score has become the most popular choice, much due to the fact that the marginal likelihood can be calculated analytically. Conveniently, this also holds for Bayesian networks with structured CPTs, such as LDAGs and CSI-trees. Using a Bayesian score with an appropriate prior, we showed through several numerical experiments on both synthetic and real-world data that the predictive properties of the inferred models can in general be improved by modeling the structure of the CPTs.

In terms of Markov networks, learning of non-chordal graphs using likelihood-based scores is very challenging and Bayesian learning has therefore been restricted to chordal graphs. We introduced the marginal pseudo-likelihood as a Bayesian alternative objective function for learning non-chordal graphs. We showed through extensive numerical experiments that the MPL, combined with an efficient search method, is competitive against recently proposed alternatives in identifying a non-chordal graph that resembles the actual graph as closely as possible. Finally, in order to obtain an analytical score function for general context-specific Markov networks, we combined the MPL with our earlier results for LDAGs. We showed that the MPL is well-justified theoretically by proving consistency of the corresponding structure estimators for both traditional and contextual Markov networks.

In future research, it would be interesting to apply more advanced search algorithms. There has lately been much research in exact learning of the graphical model structure (Bartlett & Cussens, 2013; Berg et al., 2014; Janhunen et al., 2015; Parviainen et al., 2014). In particular, exact methods developed for traditional Bayesian networks can readily be applied on Bayesian networks with structured CPTs, since the CPT structures do not impose any additional restrictions on the DAG. It would also be interesting to develop an exact method for optimizing the MPL under some additional constraints such that the scalability of the method is maintained. Another important area of future Markov network research is parameter estimation. The MPL offers a tool for high-dimensional structure learning, however, we still need to develop procedures for estimating the parameters of large-scale models. Finally, as an example of potential future applications, it would be interesting to implement and study various graphical-model-based classifiers, considering the encouraging results by Nyman et al. (2015b).

References

Bartlett, M. & Cussens, J. (2013). Advances in Bayesian network learning using integer programming. In Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence, 182–191.

Berg, J., Järvisalo, M. & Malone, B. (2014). Learning optimal bounded treewidth Bayesian networks via maximum satisfiability. In Proceedings of the 17th Conference on Artificial Intelligence and Statistics, 86–95.

Besag, J. (1975). Statistical analysis of non-lattice data. Journal of the Royal Statistical Society, Series D (The Statistician) **24**, 179–195.

Boutilier, C., Friedman, N., Goldszmidt, M. & Koller, D. (1996). Context-specific independence in Bayesian networks. In Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence, 115–123.

Buntine, W. (1991). Theory refinement on Bayesian networks. In Proceedings of the 7th Conference on Uncertainty in Artificial Intelligence, 52–60. Morgan Kaufmann.

Chickering, D. M., Heckerman, D. & Meek, C. (1997). A Bayesian approach to learning Bayesian networks with local structure. In Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence, 80–89.

Cooper, G. & Herskovitz, E. (1992). A Bayesian method for the induction of probabilistic networks from data. Machine Learning **9**, 309–347.

Corander, J. (2003). Labelled graphical models. Scandinavian Journal of Statistics **30**, 493–508.

Corander, J., Ekdahl, M. & Koski, T. (2008). Parallel interacting MCMC for learning of topologies of graphical models. Data Mining and Knowledge Discovery **17**, 431–456.

Corander, J., Gyllenberg, M. & Koski, T. (2006). Bayesian model learning based on a parallel MCMC strategy. Statistics and Computing **16**, 355–362.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L. & Spiegelhalter, D. J. (1999). *Probabilistic Networks and Expert Systems*, 1st edn. New York: Springer-Verlag.

Csiszár, I. & Talata, Z. (2006). Consistent estimation of the basic neighborhood of Markov random fields. Annals of Statistics **34**, 123–145.

Dawid, A. P. & Lauritzen, S. L. (1993). Hyper-Markov laws in the statistical analysis of decomposable graphical models. Annals of Statistics **21**, 1272–1317.

Friedman, N. & Goldszmidt, M. (1996). Learning Bayesian networks with local structure. In Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence, 252–262.

Geiger, D. & Heckerman, D. (1996). Knowledge representation and inference in similarity networks and Bayesian multinets. Artificial Intelligence **82**, 45–74.

Heckerman, D., Chickering, D. M., Meek, C., Rounthwaite, R. & Kadie, C. (2001). Dependency networks for inference, collaborative filtering, and data visualization. Journal of Machine Learning Research **1**, 49–75.

Heckerman, D., Geiger, D. & Chickering, D. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. Machine Learning **20**, 197–243.

Højsgaard, S. (2003). Split models for contingency tables. Computational Statistics & Data Analysis **42**, 621–645.

Janhunen, T., Gebser, M., Rintanen, J., Nyman, H., Pensar, J. & Corander, J. (2015). Learning discrete decomposable graphical models via constraint optimization. Statistics and Computing doi:10.1007/s11222-015-9611-4.

Koller, D. & Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

Koski, T. & Noble, J. (2009). *Bayesian networks: an introduction*. Chippenham: Wiley.

Lauritzen, S. L. (1996). *Graphical models*. Oxford: Oxford University Press.

Nyman, H., Pensar, J., Koski, T. & Corander, J. (2014). Stratified graphical models - context-specific independence in graphical models. Bayesian Analysis **9**, 883–908.

Nyman, H., Pensar, J., Koski, T. & Corander, J. (2015a). Context-specific independence in graphical log-linear models. Computational Statistics doi:10.1007/s00180-015-0606-6.

Nyman, H., Xiong, J., Pensar, J. & Corander, J. (2015b). Marginal and simultaneous predictive classification using stratified graphical models. Advances in Data Analysis and Classification doi:10.1007/s11634-015-0199-5.

Parviainen, P., Farahani, H. & Lagergren, J. (2014). Learning bounded tree-width Bayesian networks using integer linear programming. In Proceedings of the 17th Conference on Artificial Intelligence and Statistics, 751–759.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. San Francisco: Morgan Kaufmann.

Poole, D. & Zhang, N. (2003). Exploiting contextual independence in probabilistic inference. Journal of Artificial Intelligence Research **18**, 263–313.

Schwarz, G. (1978). Estimating the dimension of a model. Annals of Statistics **6**, 461–464.

Whittaker, J. (1990). *Graphical models in applied multivariate statistics*. Chichester: Wiley.

ISBN 978-952-12-3412-5